



Published in final edited form as:

Dev Cell. 2015 March 23; 32(6): 765–771. doi:10.1016/j.devcel.2015.01.013.

Pitfalls of mapping high throughput sequencing data to repetitive sequences: Piwi's genomic targets still not identified

Georgi K. Marinov^{1,*}, Jie Wang^{2,*}, Dominik Handler³, Barbara J. Wold¹, Zhiping Weng⁴, Gregory J. Hannon⁵, Alexei A. Aravin¹, Phillip D. Zamore⁶, Julius Brennecke³, and Katalin Fejes Toth¹

¹Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA

²Department of Biochemistry, University at Buffalo, Buffalo, NY 14214, USA

³Institute of Molecular Biotechnology of the Austrian Academy of Sciences IMBA, 1030 Vienna, Austria

⁴Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester MA 01605, USA

⁵Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

⁶Howard Hughes Medical Institute; RNA Therapeutics Institute and Department of Biochemistry and Molecular Pharmacology; University of Massachusetts Medical School; Worcester, MA USA

Abstract

Huang et al. (2013) recently reported that chromatin immuno-precipitation followed by sequencing (ChIP-seq) reveals the genome-wide sites of occupancy by Piwi - a piRNA-guided Argonaute protein central to transposon silencing in *Drosophila*. Their study also reported that loss of Piwi causes widespread rewiring of transcriptional patterns as evidenced by changes in RNA polymerase II occupancy across the genome. Here we reanalyze their underlying deep sequencing data and report that the data do not support the author's central conclusions.

INTRODUCTION

PIWI-clade Argonaute proteins and their small RNA guides, PIWI-interacting RNAs (piRNAs), collaborate to repress selfish genetic elements such as transposons in animal gonads (Malone and Hannon, 2009; Siomi et al., 2011). The 23–30 nt piRNAs guide PIWI proteins to targets with complementary sequences. One of the three *Drosophila* PIWI-clade proteins, Piwi is localized to the nucleus and represses transposon expression via transcriptional gene silencing (TGS). Target repression is accompanied by reduced RNA Polymerase II (Pol II) occupancy and increased tri-methylation of histone H3 Lysine 9 (H3K9me3), a mark of heterochromatin (Le Thomas et al., 2013; Rozhkov et al., 2013;

correspondence should be addressed to: kft@caltech.edu, julius.brennecke@imba.oeaw.ac.at.

*These authors contributed equally to this work

Shpiz et al., 2011; Sienski et al., 2012; Wang and Elgin, 2011). By analogy to centromeric silencing in *Schizosaccharomyces pombe* (Buhler and Moazed, 2007; Grewal, 2010), these data suggest that piRNAs guide Piwi to nascent transcripts at target loci where Piwi promotes TGS and heterochromatin formation.

Such a model is intuitively consistent with the findings of Huang et al. (Huang et al., 2013), who reported strong ChIP-seq enrichments for Piwi at many genomic regions, typically transposons, for which complementary piRNAs are observed. ChIP experiments in our own laboratories, however, have consistently failed to detect significant enrichment of Piwi at Piwi-repressed transposons, despite the use of various cross-linking conditions and different antibodies and tags for immuno-precipitation. We therefore reanalyzed the published ChIP-seq data (Huang et al., 2013). We determined (1) the degree of enrichment for Piwi at transposon loci and (2) the changes in Pol II occupancy at transposon loci upon loss of Piwi. In both cases our independent analyses failed to confirm the published conclusions. Instead, we found that different data processing methods underlie the different outcomes. We conclude that the genome-wide pattern of Piwi occupancy remains an open question despite multiple attempts to map it using contemporary ChIP-seq methods.

RESULTS

No significant enrichment of Piwi at transposon loci in the Huang et al. datasets

For the re-analysis of the Huang et al. deep sequencing data (Huang et al., 2013) we used standard read mapping procedures and retained only reads that align to the genome with 2 mismatches (for details see Supplementary Methods). For comparative purposes, we applied this strategy to a published H3K9me3 ChIP-seq dataset from *Drosophila* ovaries (Muerdter et al., 2013). This histone mark is enriched in heterochromatin and on transposons and other genomic repeats. It is also present at transposon insertions repressed by nuclear Piwi via the piRNA pathway.

To ask whether the Piwi ChIP-seq dataset was enriched for transposon sequences, we first mapped all genome-mapping ChIP-seq and input control reads to a comprehensive list of consensus transposon sequences for *Drosophila melanogaster*. For each, we calculated normalized RPM values (Reads Per Million sequenced reads; for details see Supplementary Methods). This resulted in Piwi occupancy levels for transposons that were indistinguishable from background (Fig. 1A). In contrast, the H3K9me3 mark was as much as tenfold enriched over most transposons. These results are in marked contrast with the conclusion that “~86% of the Piwi ChIP-seq signal overlaps with transposons and repetitive sequences” (Huang et al., 2013). Our analysis of the Piwi ChIP-seq data does also not support the ChIP-qPCR data presented by Huang et al. in their Fig. S1 (Huang et al., 2013), which shows that DNA fragments of two transposons (*F*-element and *I360*) were retrieved at least tenfold more efficiently in Piwi ChIP experiments compared to control IPs; in fact, neither of these transposons was detectably enriched in the Piwi ChIP-seq dataset in our analysis, although both were significantly enriched in the H3K9me3 ChIP-seq data (Fig. S1). The positive H3K9me3 ChIP-seq outcome from our analysis shows that a heterochromatin-associated mark can be and was successfully captured and associated DNA efficiently sequenced. This argues against scenarios in which ChIP-enriched heterochromatic regions are detected by

qPCR, even though they are missed by ChIP-seq because they are especially poor substrates for library building and/or sequencing. Of note, Huang et al. reported similar Piwi enrichments when ChIP-qPCR experiments were conducted from dissected ovaries compared to whole flies (Fig S1 of Huang et al., 2013). As Piwi is expressed at high levels only in gonadal cells, ChIP-qPCR signals are predicted to be even diluted by somatic nuclei when gonads are compared with whole flies.

Next, we analyzed the Piwi ChIP-seq data at the genomic level. Figure 1B depicts a genomic region harboring three transposon insertions; this same region is shown in Fig. 2C of Huang et al. (Huang et al., 2013). Read coverage for Piwi ChIP-seq and the corresponding input datasets was calculated three ways: (1) considering only reads that map the genome uniquely; (2) considering all reads mapping to the genome but normalizing each for the number of times it mapped to the genome; and (3) considering all genome matching reads without any normalization. None of the three transposon insertions nor their immediate genomic neighborhood stood out in the Piwi ChIP data compared to the background when (1) unique reads or (2) normalized reads were considered (Fig. 1B). When the reads were (3) not corrected for mapping to multiple genomic sites, transposons emerged as strong peaks relative to flanking genomic sequences. However, transposons also emerged as strong peaks when the control dataset, the input genomic DNA itself, was mapped without accounting for mapping multiplicity. We used each of the three mapping strategies to determine the genome-wide average read density for Piwi ChIP and input datasets over the three major transposable element classes in *Drosophila* (e.g., LINE elements in Fig. 2). In all cases, we found no enrichment of Piwi over background, whereas the H3K9me3 dataset again displayed strong enrichment.

Finally, we asked whether the Piwi ChIP dataset was enriched for Piwi occupying to a subset of the thousands of transposon insertions in the *Drosophila* genome. Such a subset might go undetected when analyzing genome-wide average signals. We compared the enrichment of Piwi at individual transposons with that of eleven transcription factors whose genome-wide occupancy has been determined from early fly embryos (modENCODE Consortium, 2010; Negre et al., 2011); none of these developmental regulators is expected to be selectively enriched at transposon loci. Again, we found no specific enrichment of Piwi at transposon loci: the enrichment of Piwi at transposons was well within the range of enrichment observed for the transcription factors on the same set of transposons (Fig. 1C). In contrast, the H3K9me3 mark was strongly enriched over all transposon classes. Taken together, these analyses show that the published Piwi ChIP-seq datasets do not support a specific enrichment of Piwi at transposons.

The Huang et al. computational pipeline generates artificial enrichment of ChIP-seq datasets at repetitive loci

To identify the discrepancy between our standard analysis pipeline and that of Huang et al., we examined the computational pipeline used in their studies (originally described in (Yin et al., 2011), which the authors kindly shared with us. Rather than defining enrichments by the ratio of ChIP versus input sample reads, the Huang et al. pipeline identifies genomic regions of Piwi enrichment via a multistep procedure (see Fig. S2 and Supplementary Methods for

details). Two features of this pipeline could artificially amplify minor differences between ChIP and control datasets into large apparent enrichments at transposons. First, the pipeline makes no correction for reads mapping to multiple genomic locations. Of course, one single read must come from a single genomic locus, no matter how many times it maps to the genome, so all widely used mapping software either randomly assign a multiply mapping read to a single locus or apportion the read among the multiple loci. Without such standard corrections for mapping multiplicity, all datasets—both ChIP-seq and input genomic DNA—produce artificially elevated signals at repetitive loci such as transposons. Considering that they apply a cutoff threshold (see Experimental Procedures), this artificially elevated signal focuses the analysis strongly towards repetitive regions. Because Huang et al. also allowed 5 mismatches and 4 indels when mapping their reads to the genome, they inflate the number of reads from related loci with high sequence divergence: i.e., transposons and repetitive sequences. Second, although the subsequent analysis does take the input datasets into account, it does so in a non-standard way by applying nonlinear transformations to the resulting signal tracks. The consequence of this nonlinear approach is that the final score displays positive enrichments but sets negative enrichments (i.e. depletions) to zero. Ultimately, the combination of these steps leads to exclusively positive enrichments preferentially at transposons (Fig. 1B), while signal in the direction of depletion is obscured. The algorithm is particularly prone to create artificial peaks from ChIP-seq datasets with low signal-to-noise ratios (see below).

By way of example, we recapitulated the Huang et al. analysis, but swapping the input background and Piwi ChIP-seq data, and then calculated the percentage of ‘signal’ at annotated repeats. Strikingly, treating the genomic DNA input as the experiment and the PIWI ChIP-seq as the control produced strong signal enrichment at transposons. In fact, an even higher proportion of the final signal mapped to repeats in this nonsensical analysis than when the data sets were correctly assigned to experiment and control (Fig. 3A). The identity of the particular repeats contributing to the final signal, however, differed as is expected if the result stems from mistaking amplified positive noise for signals. Figure 3B displays the final Huang et al. scores for Piwi ChIP-seq over background and background over Piwi ChIP-seq at three individual, full-length transposon insertions (Fig. 3B). While some transposon insertions showed high signal in the Piwi/background track (e.g. *roo*), others showed high “enrichment” in the background/Piwi track (e.g. *Max*) and some transposon insertions showed a “mixed” signal, in which different portions of the element are highly “enriched” in either the background or the ChIP tracks (e.g. *blood*). These observations also suggest that the Huang et al. pipeline has the somewhat counterintuitive effect of generating much higher enrichments over transposons for ChIP datasets that contain very little or no true signal than it does for ChIP datasets that are strongly enriched at genomic features other than transposons. In the latter case, transposons are globally depleted relative to the control because a high fraction of reads is concentrated in regions of true occupancy located elsewhere in the genome. This is not the case in input and poorly enriching ChIP experiments leading to a higher apparent enrichment over TE sequences. Indeed, when we calculated the percentage of signal at transposons for the modENCODE transcription factor ChIP-seq dataset using the method of Huang et al., we observed highly variable results (Fig.

3C). For some developmental regulators, the Huang et al., signal on repeats was similar to the Piwi dataset, while other factors displayed little signal on transposons.

The experimental characterization of the true genomic distribution of Piwi on chromatin thus remains an unresolved challenge. The difficulty in obtaining Piwi ChIP-seq signal likely reflects the added complexity of obtaining DNA sequences transiently tethered to Piwi protein via nascent RNA. The inherent difficulty in shearing heterochromatin may also contribute to the problem (Teytelman et al., 2009).

No support for widespread transcriptional changes in *piwi* mutants

Based on the same computational pipeline, Huang et al. also reported that in *piwi* mutants Pol II is broadly redistributed from protein-coding genes to transposons. We calculated consensus transposon RPM values for the Pol II ChIP-seq datasets and their respective controls (Fig. 4A). We found no clear differences between Pol II enrichments over transposons in wild-type versus *piwi* mutant flies. In both samples, Pol II was depleted at transposons compared to the input (Fig 4A and B), likely due to its enrichment at protein-coding genes in the Pol II ChIP-seq data but not the input control. In contrast, Huang et al. reported that Pol II concentrated on transposons in *piwi* mutants compared to wild type. A meta-profile of Pol II occupancy at all protein-coding loci showed a ~2-fold greater enrichment at promoters in wild type compared to the mutant (Fig. S3). For the *piwi* mutant data set this means that proportionally fewer reads originate from expressed genes versus the remainder of the genome. In consequence, more background reads from transposons are recovered, and these are then amplified by the Huang et al. pipeline.

Taken together, our analyses find no support for a widespread role of Piwi in specifying patterns of transcription at transposons in the published datasets. On the other hand, loss of Piwi has been shown in several studies to lead to pronounced changes in Pol II occupancy at piRNA-pathway-repressed transposon loci (Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012). We note that these studies analyzed isolated ovaries or cultured ovarian somatic cells rather than entire flies. One conclusion of these studies is that biologically meaningful analyses of Piwi function using ChIP experiments require the use of isolated tissues where nuclear Piwi is highly expressed: the gonads.

The biologically relevant pattern of Piwi genomic occupancy remains unknown. Piwi associates with piRNAs complementary to virtually all transposon families, and loss of Piwi leads to the selective loss of the H3K9me3 mark at several transposon insertions (Sienski et al., 2012). These observations suggest that sequence complementarity between piRNAs and nascent target transcripts dictate the chromatin occupancy of Piwi. Considering the technical difficulties that have surrounded Piwi ChIP-seq, a first step towards identifying Piwi binding sites should be to verify direct occupancy at one or a few functional genomic target sites using alternative methods such as Dam-ID (van Steensel and Henikoff, 2000). These validated sites could then be used as internal standards to establish approaches for the mapping of Piwi on chromatin across the genome.

EXPERIMENTAL PROCEDURES

Data processing

A detailed description of our computational analysis is in the Supplementary Materials section. In summary, the data from Huang et al. (2013) as well as from Muerdter et al. (2013) were processed using both the Huang et al. pipeline and more conventional approaches incorporating three different signal normalization approaches. We aligned reads to the *Drosophila melanogaster* genome (dm3) using Bowtie (Langmead, 2010; version 0.12.7) and then generated signal tracks by calculating: (1) normalized (RPM, Reads Per Million mapped reads) coverage using only uniquely alignable reads; (2) RPM coverage using all alignments, weighting each according to the number of locations in the genome to which the read maps; and (3) RPM coverage using all alignments treated as if they were uniquely aligned reads (i.e., without normalization for multi-mappers, as in the Huang et al. pipeline).

The Huang et al. pipeline was reproduced according to the description and parameters presented in Yin et al. (2011). Briefly, it begins by recursively aligning reads with SOAP, allowing up to 5 mismatches and 4 indels. Alignments are then converted into 5' coordinates, the chromosomes are split into 50 bp bins, and each alignment contributes to 10 bins according to a weighting scheme that decreases its weight in more distant bins. The scores are then normalized according to the total number of alignments (rather than the total number of reads, i.e. no multi-mapping normalization is applied) and a “critical value” is calculated for each ChIP/Input pair such that beyond that value the bin values are always higher in the ChIP than in the control dataset (Fig. S2); a normalizer score is calculated based on the bins with values lower than the critical value, and is applied to the ChIP. The ChIP is further normalized by subtracting the background. Critically, when this step is performed, negative values are set to zero, leading to loss of data over regions of depletion relative to background. Finally, scores are divided by the trimmed mean, log-transformed, and again set to zero if negative.

Repeat analysis

RepeatMasker annotation, downloaded from the UCSC Genome Browser, was used for the analysis of repetitive element coverage in genomic space. Consensus repetitive elements were downloaded from FlyBase (Marygold et al., 2013); reads were aligned against them using Bowtie, allowing for 3 mismatches and unlimited multi-mappers, and normalized RPM values calculated for each element.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Buhler M, Moazed D. Transcription and RNAi in heterochromatic gene silencing. *Nature structural & molecular biology*. 2007; 14:1041–1048.
- Grewal SI. RNAi-dependent formation of heterochromatin and its diverse functions. *Current opinion in genetics & development*. 2010; 20:134–141. [PubMed: 20207534]

- Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H. A major epigenetic programming mechanism guided by piRNAs. *Developmental cell*. 2013; 24:502–516. [PubMed: 23434410]
- Langmead, B. Aligning short sequencing reads with Bowtie. In: Baxevanis, Andreas D., et al., editors. *Current protocols in bioinformatics / editorial board*. Vol. Chapter 11. 2010. p. 17
- Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Toth KF. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes & development*. 2013; 27:390–399. [PubMed: 23392610]
- Malone CD, Hannon GJ. Small RNAs as guardians of the genome. *Cell*. 2009; 136:656–668. [PubMed: 19239887]
- Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ. FlyBase, c. FlyBase: improvements to the bibliography. *Nucleic acids research*. 2013; 41:D751–757. [PubMed: 23125371]
- Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. modENCODE. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*. 2010; 330:1787–1797. [PubMed: 21177974]
- Muerdter F, Guzzardo PM, Gillis J, Luo Y, Yu Y, Chen C, Fekete R, Hannon GJ. A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*. *Molecular cell*. 2013; 50:736–748. [PubMed: 23665228]
- Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al. A cis-regulatory map of the *Drosophila* genome. *Nature*. 2011; 471:527–531. [PubMed: 21430782]
- Rozhkov NV, Hammell M, Hannon GJ. Multiple roles for Piwi in silencing *Drosophila* transposons. *Genes & development*. 2013; 27:400–412. [PubMed: 23392609]
- Shpiz S, Olovnikov I, Sergeeva A, Lavrov S, Abramov Y, Savitsky M, Kalmykova A. Mechanism of the piRNA-mediated silencing of *Drosophila* telomeric retrotransposons. *Nucleic acids research*. 2011; 39:8703–8711. [PubMed: 21764773]
- Sienski G, Donertas D, Brennecke J. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*. 2012; 151:964–980. [PubMed: 23159368]
- Siomi MC, Sato K, Pezic D, Aravin AA. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol*. 2011; 12:246–258. [PubMed: 21427766]
- Teytelman L, Ozaydin B, Zill O, Lefrancois P, Snyder M, Rine J, Eisen MB. Impact of chromatin structures on DNA processing for genomic analyses. *PloS one*. 2009; 4:e6700. [PubMed: 19693276]
- van Steensel B, Henikoff S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nature biotechnology*. 2000; 18:424–428.
- Wang SH, Elgin SC. *Drosophila* Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108:21164–21169. [PubMed: 22160707]
- Yin H, Sweeney S, Raha D, Snyder M, Lin H. A high-resolution whole-genome map of key chromatin modifications in the adult *Drosophila melanogaster*. *PLoS genetics*. 2011; 7:e1002380. [PubMed: 22194694]

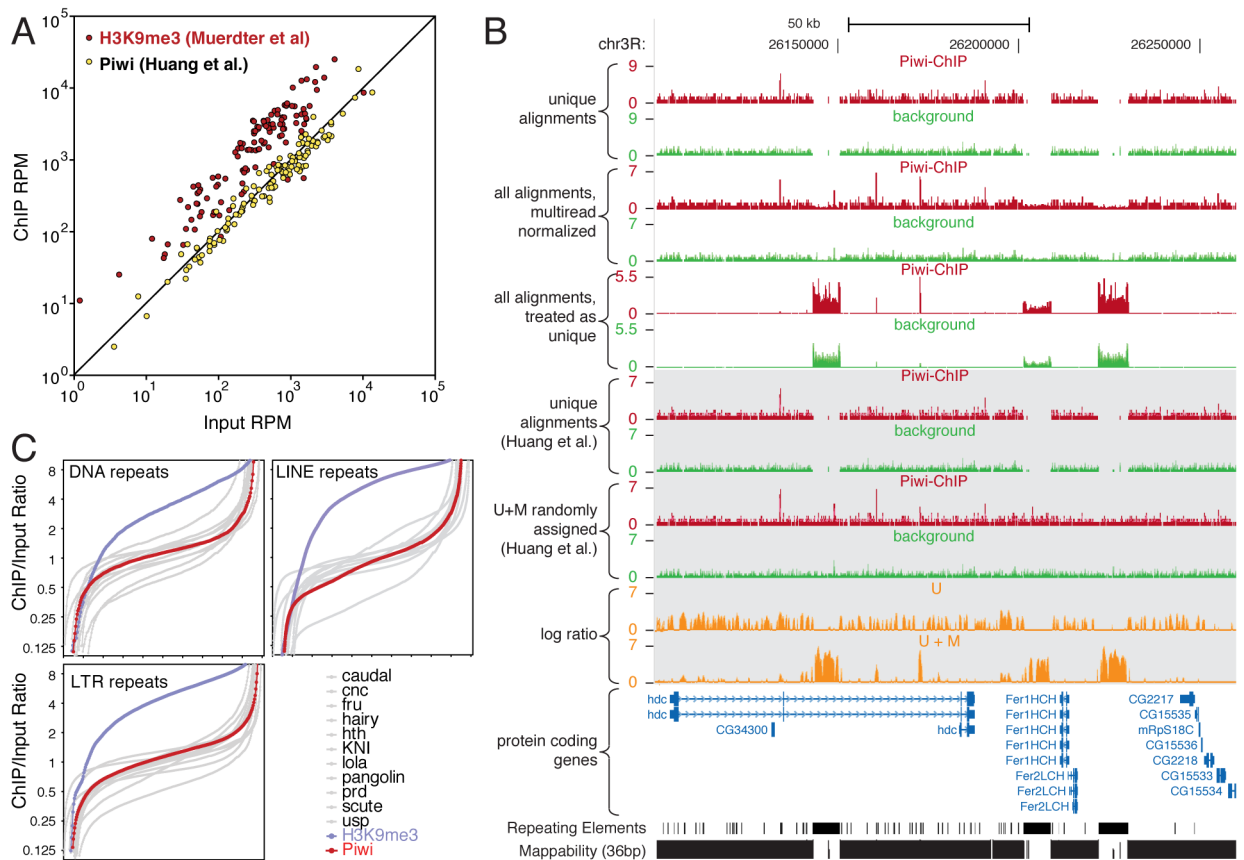


Figure 1. Piwi is not enriched over transposons in the Huang et al. dataset

(A) Absence of enrichment in the Piwi ChIP-seq dataset and high enrichment of H3K9me3 (from Muerdter et al., 2013) over consensus transposons; each dot corresponds to a transposon consensus sequence. (B) The concentration of Piwi signal over transposons in the Huang et al. dataset arises from failure to normalize multiply mapping- reads. Shown is the region from Fig. 2C of Huang et al. (2013). Top: Piwi ChIP-seq and background (input) data from Huang et al. showing: (1) unique alignments; (2) all alignments, with reads normalized for mapping multiplicity; and (3) all alignments, with all reads treated as unique. Bottom: data processed per Huang et al. The enrichment of Piwi over repetitive elements is only observed when no multi-read normalization is applied and is seen in both ChIP and control datasets. (C) The minimal Piwi ChIP-seq enrichment observed over some individual transposable elements is well within the range of experimental noise. Shown is the cumulative distribution function (CDF) of the ratio between total ChIP RPM and control/background RPM for each DNA, LINE or LTR repetitive element (each dot represents an individual TE insertion). Piwi ChIP-seq data from Huang et al. (red) and H3K9me3 data from Muerdter et al. (blue) are plotted alongside the cumulative distribution for 11 transcription factor ChIP-seq datasets from modENCODE (gray), for which there is no expectation of enrichment at repetitive elements. Only repeat instances with at least 10 RPM in at least one of the ChIP and control datasets for each ChIP/background pairing were included. H3K9me3 showed high average enrichment over background at most of the

elements in all three classes. In contrast, the Piwi ChIP-seq data was well within the range of the distributions for modENCODE transcription factors.

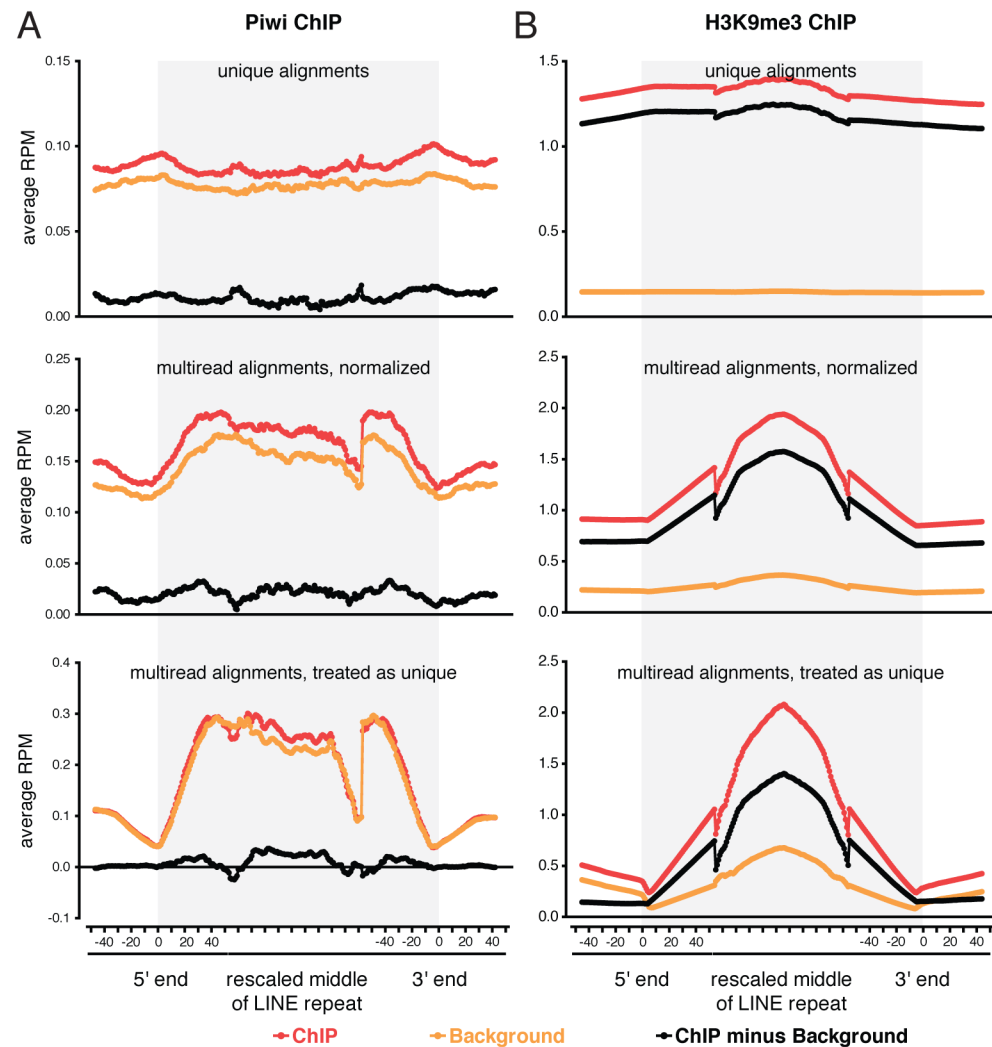


Figure 2. Distribution of Piwi and H3K9me3 over repetitive elements in the genome
 Shown is the average signal distribution over LINE repetitive elements for ChIP (red) and background (yellow) datasets for Piwi from Huang et al. (2013) (A) and for H3K9me3 from Muerdter et al. (2013) (B). The background-normalized enrichment is in black. The 100 bp around the beginning and the end of individual elements are shown to scale; the rest of each LINE element is rescaled to 100 units. The repeat-Masker repetitive element annotation from the UCSC Genome Browser was used. A clear enrichment over background is observed in H3K9me3 datasets, even when only uniquely aligning reads are considered. In contrast, the Piwi dataset from Huang et al. is essentially indistinguishable from background.

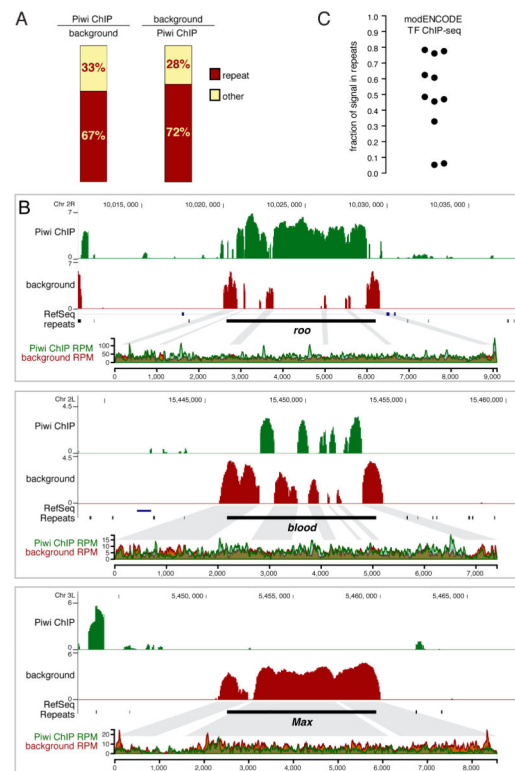


Figure 3. The Huang et al. data processing pipeline generates artificial enrichment over repetitive regions

The Piwi ChIP-seq and input/background datasets were processed following the Huang et al. pipeline (“Piwi ChIP”). In addition, the pipeline was also run swapping the ChIP and the input, i.e. the control sample was treated as ChIP and vice versa, resulting in the “background” track. (A) The fraction of signal mapping to transposable elements was calculated, revealing higher “enrichment” in the background than in the Piwi ChIP-seq data set. (B) Strong apparent enrichment over individual transposable elements was observed in the ChIP track (upper track), as reported by Huang et al., but also in the background track (lower track), and even over different portions of the same transposable element in both tracks (middle track), strongly arguing that the enrichment over transposable elements reported by Huang et al. is a computational artifact. Signal observed on individual copies correlates well with enrichment profiles when mapped to the consensus sequence of the respective transposons (shown below each track). Sequences showing “enrichment” in the background are indicated with gray blocks to depict the correlations between the signal on individual TE copies and the consensus sequence. (C) Fraction of signal (calculated with the Huang et al. pipeline) mapping to transposable elements for the modENCODE transcription factor set.

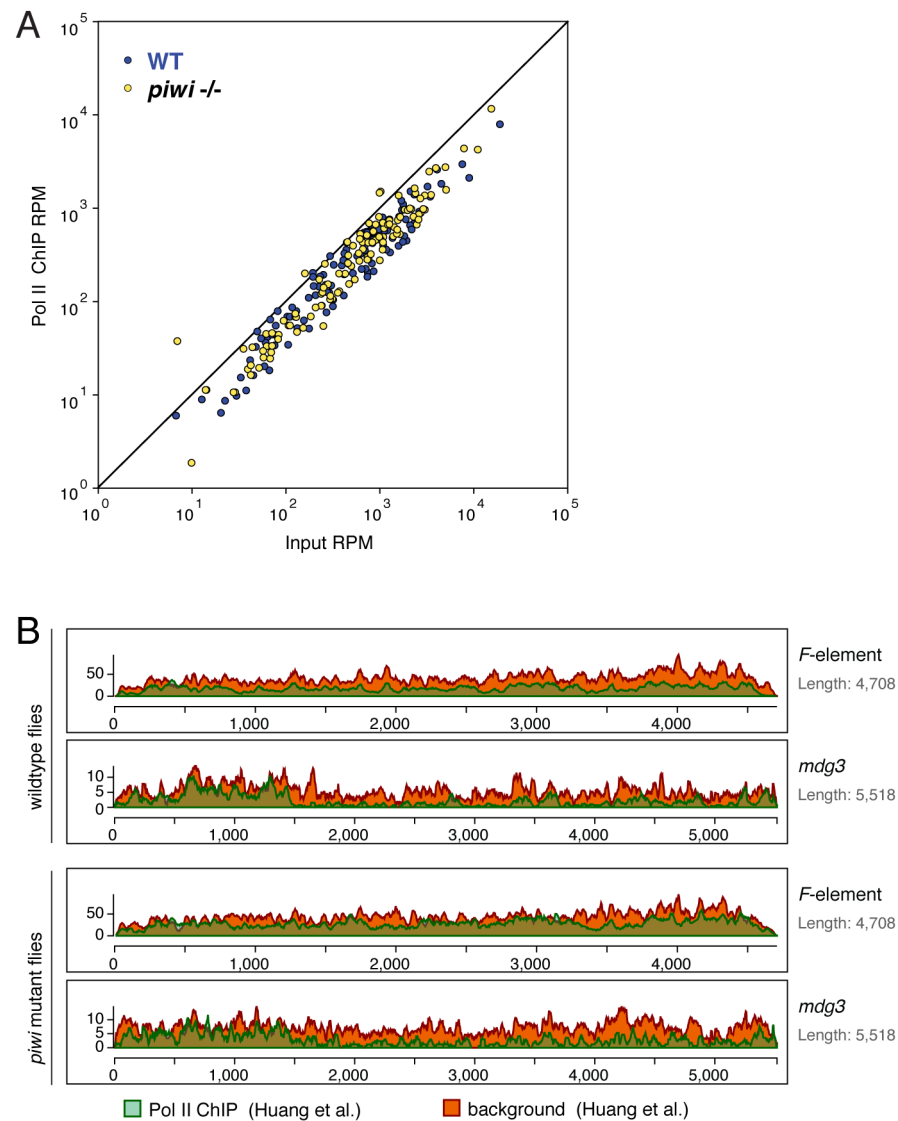


Figure 4. No redistribution of Pol II over transposons is observed in *piwi* mutant files
 (A) Scatter plot displaying Pol II ChIP-seq RPM values versus input RPM values over consensus transposable elements in wild type and *piwi* mutant flies. (B) Shown are Pol II ChIP-seq and input RPM levels over the transposon consensus sequences of *F*-element and *mdg3*.